UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
09/995,087	11/27/2001	Paul Michael Dantzig	YOR920010320US1	9571

7590 10/02/2008

Kin-Wah Tong, Esq.
Moser, Patterson & Sheridan, LLP
595 Shrewsbury Avenue - Suite 100
Shrewsbury, NJ 07702

EXAMINER	
NAWAZ, ASAD M	

ART UNIT	PAPER NUMBER
2155	

MAIL DATE	DELIVERY MODE
10/02/2008	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Office Action Summary	Application No.	Applicant(s)	
	09/995,087	DANTZIG ET AL.	
	Examiner	Art Unit	
	ASAD M. NAWAZ	2155	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE *3* MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *10 July 2008*.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1-5,7-10 and 32-40* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1-5,7-10 and 32-40* is/are rejected.

7)☒ Claim(s) *10* is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on *27 November 2001* is/are: a)☒ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☐ Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date _____.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____ .
5) ☐ Notice of Informal Patent Application
6) ☐ Other: _____.

DETAILED ACTION

1. This action is responsive to the RCE received 7/18/08. Claims 1, 3-4, 7-10, 32-35, and 37-40 have been amended. Claims 11, 14-15, 17, 20-23, 26-28, 31, and 41-12 have been canceled. No other claims have been amended, added, or canceled. Accordingly, claims 1-5, 7-10, 32-40 are pending.

Continued Examination Under 37 CFR 1.114

2. A request for continued examination under 37 CFR 1.114, including the fee set forth in 37 CFR 1.17(e), was filed in this application after final rejection. Since this application is eligible for continued examination under 37 CFR 1.114, and the fee set forth in 37 CFR 1.17(e) has been timely paid, the finality of the previous Office action has been withdrawn pursuant to 37 CFR 1.114. Applicant's submission filed on 07/18/08 has been entered.

Response to Arguments

3. Applicant's arguments with respect to claims 1-5, 7-10 and 32-40 have been considered but are moot in view of the new ground(s) of rejection.

Claim Objections

4. Claim 10 is objected to because of the following informalities: Claim 10 does not end in a period. Appropriate correction is required.

Claim Rejections - 35 USC § 103

5. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

6. Claims 1-3, 5, 7-10, 32-34, and 36-40 are rejected under 35 U.S.C. 103(a) as being unpatentable over Krishnan (US Pub No 2005/0243862) further in view of Stark et al (US Pub No 2003/0149735) hereinafter referred to as Stark.

As to claim 1, Kirshnan teaches a method comprising the steps of: determining a load on said primary server (fig 5, 0085, 0087, bandwidth utilization is measured for a period of time and repeated in intervals);

if the load on said primary server is less than a first threshold, serving processing requests at said primary server (fig 5, 0040, 0044-0045, 0077; if the load is less than a first threshold (within a first zone) no actions are taken and the requests are carried out normally);

only if the load on said primary server exceeds said first threshold (0045, 0077; if load exceeds a first threshold but is less than a second threshold (second zone), first set of throttling actions are taken).

and if the load on said primary server exceeds a second threshold, throttling at least one of said processing requests (Fig 5, 0045-0046, 0077; if the load exceeds the second threshold a second set of throttling actions are taken).

However, Krishnan does not explicitly indicate offloading at least a portion of said processing requests to any one of said plurality of offload servers, wherein all of said plurality of offload servers are configured to process said processing requests and the offloaded at least a portion of said processing requests is the only work handled by said plurality of offload servers.

Stark teaches a method, in a network comprising a primary server and a plurality of offload servers, for dynamic offloading of processing requests from said primary server to any one of said plurality of offload servers (0007, 0022, 0051; requests are offloaded to spare/standby redundant secondary servers in case of failover)

offloading at least a portion of said processing requests to any one of said plurality of offload servers (0051; requests are offloaded to spare/standby redundant secondary servers in case of failover/switchover),

wherein all of said plurality of offload servers are configured to process said processing requests (0022 0048, highly availably redundant servers provide a cluster of offload servers that are master-eligible and have the capability to act as a master node)

and the offloaded at least a portion of said processing requests is the only work handled by said plurality of offload servers (0048, 0063; the secondary servers are in a spare or offline state and are thus not processing any work).

It would have been obvious to one of ordinary skill in the art at the time of the invention to incorporate the teachings of Stark into those of Krishnan to make the system more efficient. Allowing dedicated backup servers to assist in easing the load would help serve content quicker. Krishnan proposes using a plurality of servers (0002)

from one machine on separate servers (0053) (like Stark) and implementing the BT (bandwidth throttling) system as a separate management component (0053). Therefore this "allows both that server as well as other servers to perform more efficiently" and thus minimize the impact on various network services (0033, 0048).

As to claim 2, Krishnan teaches the method of claim 1 wherein said load comprises bandwidth utilization and said first threshold is a network bandwidth utilization of said primary server (abstract, 0078; the threshold can be a factor of processor, memory, or bandwidth utilization).

As to claim 3, Krishnan teaches the method of claim 1 wherein the said load comprises CPU utilization and said first threshold is a central processing unit (CPU) utilization of said primary server (0078; the threshold can be a factor of processor, memory, or bandwidth utilization).

As to claim 5, Krishnan teaches the method of claim 1 comprising an incoming Web request (0049, 0052; web services and email services can be handled by an ISP server).

However, Krishnan does not teach offloading processing requests includes routing an incoming request to a selected offload server.

Stark teaches offloading processing requests includes routing an incoming request to a selected offload server (0116, 0122; during shutdown/failure, a gradual process offloads workload to the secondary server (primary server can still be recovered in this state)).

It would have been obvious to one of ordinary skill in the art at the time of the invention to incorporate the teachings of Stark into those of Krishnan to make the system more efficient. Allowing dedicated backup servers to assist in easing the load would help serve content quicker. This "allows both that server as well as other servers to perform more efficiently (0033)" and "the impact on various network services are minimized" (0048).

As to claim 7, Krishnan teaches the method of claim 1 wherein throttling at least one of said processing request requests includes returning a page to a user indicating that a server is overloaded (0035; an alternative interprocess communication scheme enables the server to notify the client that the server is busy).

As to claim 8, Krishnan teaches the method of claim 1 wherein throttling at least one of said processing requests includes dropping the at least one of said processing requests without returning any information to a user (0035, when a client transmits a request, the only communication is the response to request without notification to user about server status).

As to claim 9, Krishnan teaches the method of claim 1 wherein throttling at least one of said processing requests includes returning a page to a user indicating that a server is overloaded if said load exceeds said second threshold, and dropping said at least one of said processing request requests if said load exceeds a third threshold (0035, further escalation includes the rejection of large write and/or large transmit operations during a severe overload condition).

As to claim 10, Krishnan teaches the method of claim 1 wherein the a determination of which of said plurality of offload servers that at least a portion of said processing requests is to be offloaded to is based on one or more of a group including: a client identity, a client gateway (Internet Protocol) address, a price of the offload service, or a current or previous load on the any one of said plurality of offload servers (0016, 0037, delays and actions are performed based on class of services to which a user belongs).

As to claim 32 Krishnan teaches a method for allocating processing requirements on an Internet Protocol network (0072) comprising:

periodically evaluating processing requests to determine a load on said primary server and determine if said load exceeds a first threshold for a predetermined period of time (fig 5, 0036, 0085, 0087, bandwidth utilization is measured for a period of time and repeated in intervals) ;

only if said load does not exceed said first threshold, directing said processing requests to said primary server (fig 5, 0044-0045, 0077; if the load is less than a first threshold (within a first zone) no actions are taken and the requests are carried out normally);

and if the load on said primary server exceeds a second threshold, throttling at least one of said processing requests (Fig 5, 0045-0046, 0077; if the load exceeds the second threshold a second set of throttling actions are taken).

However, Krishnan does not explicitly indicate directing at least one processing request to any one of said plurality of offload servers, wherein all of said plurality of

offload servers are configured to process said processing request and the offloaded at least a portion of said processing requests is the only work handled by said plurality of offload servers.

Stark teaches directing at least one processing request to any one of said plurality of offload servers (0051; requests are offloaded to spare/standby redundant secondary servers in case of failover/switchover),

wherein all of said plurality of offload servers are configured to process said processing request (0022 0048, highly availably redundant servers provide a cluster of offload servers that are master-eligible and have the capability to act as a master node);

and the offloaded at least a portion of said processing requests is the only work handled by said plurality of offload servers (0048, 0063; the secondary servers are in a spare or offline state and are thus not processing any work).

It would have been obvious to one of ordinary skill in the art at the time of the invention to incorporate the teachings of Stark into those of Krishnan to make the system more efficient. Allowing dedicated backup servers to assist in easing the load would help serve content quicker. Krishnan proposes using a plurality of servers (0002) from one machine on separate servers (0053) (like Stark) and implementing the BT (bandwidth throttling) system as a separate management component (0053). Therefore this “allows both that server as well as other servers to perform more efficiently (0033)” and “the impact on various network services are minimized” (0048).

As to claim 33, Krishnan teaches the method of claim 32 wherein said load comprises network bandwidth and said first threshold is a measure of the network

bandwidth utilization of said primary server (abstract, 0078; the threshold can be a factor of processor, memory, or bandwidth utilization).

As to claim 34, Krishnan teaches the method of claim 32 wherein said load comprises central processing unit (CPU) utilization and said first threshold is a measure of the CPU utilization of said primary server (abstract, 0078; the threshold can be a factor of processor, memory, or bandwidth utilization).

As to claim 36, Krishnan teaches the method of claim 32 comprising an incoming Web request (0049, 0052; web services and email services can be handled by an ISP server).

However, Krishnan does not teach offloading processing requests includes routing an incoming request to a selected offload server.

Stark teaches offloading processing requests includes routing an incoming request to a selected offload server (0116, 0122; during shutdown/failure, a gradual process offloads workload to the secondary server (primary server can still be recovered in this state)).

It would have been obvious to one of ordinary skill in the art at the time of the invention to incorporate the teachings of Stark into those of Krishnan to make the system more efficient. Allowing dedicated backup servers to assist in easing the load would help serve content quicker. This “allows both that server as well as other servers to perform more efficiently” and thus minimize the impact on various network services are minimized (0033, 0048).

As to claim 37, Krishnan teaches the method of claim 32 wherein said throttling at least one of said processing requests comprises returning a page to a user indicating that a server is overloaded (0035; an alternative interprocess communication scheme enables the server to notify the client that the server is busy).

As to claim 38, Krishnan teaches the method of claim 32 wherein said throttling of at least one of said processing requests comprises dropping the at least one of said processing requests without returning any information to a user (0035, when a client transmits a request, the only communication is the response to request without notification to user about server status).

As to claim 39 Krishnan teaches the method of claim 32 wherein the throttling of at least one of said processing request requests comprises by returning a page to a user indicating that the primary server is overloaded if the primary server load exceeds the second threshold, and further comprising dropping the at least one of said processing requests if the load exceeds a third threshold (0035, further escalation includes the rejection of large write and/or large transmit operations during a severe overload condition).

As to claim 40, Krishnan teaches the method of claim 32 further including determining which of said plurality of offload servers that said at least one of said processing request requests is to be offloaded to is based on one or more of a group including: a client identity, a client gateway (Internet Protocol) address, a price of the offload service, or a current or previous load on the at least one any one of said plurality

of offload server servers (0016, 0037, delays and actions are performed based on class of services to which a user belongs).

7. Claims 4 and 35 are rejected under 35 U.S.C. 103(a) as being unpatentable over Krishnan and Stark as applied to claims 1-3, 5, 7-10, 32-34, and 36-40 above, and further in view of Nepustil (US Patent No 6,240,454).

As to claim 4, Krishnan and Stark teach the method of claim 1, however does not explicitly indicate serving the processing requests at said primary server includes returning a page to a user wherein all the embedded objects in the page have links to said primary server; and offloading at least a portion of the processing request to any one of said servers includes serving a base page at said primary server in which the links for embedded objects point to any one of said plurality of servers

Nepustil teaches the method of claim 1 wherein serving the processing requests at said primary server includes returning a page to a user wherein all the embedded objects in the page have links to said primary server; and offloading at least a portion of the processing request to any one of said servers includes serving a base page at said primary server in which the links for embedded objects point to any one of said plurality of servers (fig 2, col 2, lines 47-67, col 3, lines 50- col 4, line 16 and 32-49).

It would have been obvious to one of ordinary skill in the art at the time of the invention to incorporate the teachings of Nepustil into those of Krishnan and Stark to make the system more robust. By dispersing the system and its resources, one ensures that the system does not have a single point of failure, as well as being

hardened against total failure. A quick response time is also received as workload is done in parallel.

As to claim 35, Krishnan and Stark teach the method of claim 32 however they do not explicitly indicate directing said processing requests to said primary server further includes returning a page to a user wherein all the embedded objects in the page have links to said primary server and directing at least one processing request to any one of said plurality of offload servers further includes serving a base page at said primary server in which the links for embedded objects point to said any one of said plurality of offload servers.

Nepustil teaches the method of claim 1 wherein serving the processing requests at said primary server includes returning a page to a user wherein all the embedded objects in the page have links to said primary server; and offloading at least a portion of the processing request to any one of said servers includes serving a base page at said primary server in which the links for embedded objects point to any one of said plurality of servers (fig 2, col 2, lines 47-67, col 3, lines 50- col 4, line 16 and 32-49).

It would have been obvious to one of ordinary skill in the art at the time of the invention to incorporate the teachings of Nepustil into those of Krishnan and Stark to make the system more robust. By dispersing the system and its resources, one ensures that the system does not have a single point of failure, as well as being hardened against total failure. A quick response time is also received as workload is done in parallel.

Prior Art

8. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure.

A) Rumsewics et al, US Patent No. 6,832,255, directed towards a method and apparatus for efficient access control of resources.

Conclusion

9. Any inquiry concerning this communication or earlier communications from the examiner should be directed to Asad M. Nawaz whose telephone number is (571) 272-3988. The examiner can normally be reached on M-F 8-4:30.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Saleh Najjar can be reached on (571) 272-4006. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free).

/Asad M Nawaz/
Examiner, Art Unit 2155